

Různé metody manažerství kvality

Práce č. 14: Výběr ze základního souboru – intervalový a bodový odhad pravděpodobnosti

Obsah

Obsah.....	2
1 Úvod.....	3
1.1 Náhodný výběr.....	3
1.2 Různé druhy výběru.....	4
1.3 Vliv počtu vybraných kusů.....	5
2 Bodový a intervalový odhad pravděpodobnosti	9
2.1 Intervalový odhad pravděpodobnosti	10
2.2 Bodový odhad - výpočet pravděpodobnosti konkrétní hodnoty	14
Použitá literatura	17

1 Úvod

V dnešním díle si probereme problematiku výběru ze základního souboru a vliv počtu vybraných kusů. Následně intervalový odhad a bodový odhad pravděpodobnosti.



1.1 Náhodný výběr

Náhodný výběr by měl splňovat:

Pravdivost: neodstraňovat odlehlé hodnoty, hodnoty mimo limit anebo blízko limitu, pokud nedokážeme zdůvodnit:

- jejich neirelevantnost – např. nesériová výroba, při které technik/ technolog/ kvalita nastavovali parametry anebo něco zkoušeli;
- co je způsobilo anebo co se za nimi „skrývá“
 - Kolem šel vozík a vibrace narušily měření – jedná se tedy o narušení měřicího systému, ne o hodnotu s odlišnými parametry.
 - První kus/-y – než se proces stabilizuje – je důležité, co děláme s prvními kusy, pokud je standardně vyhazujeme (tedy „neměli by se“ dostat k zákazníkovi), můžeme je odstranit z analyzovaného souboru. Ale pokud takové kusy se k zákazníkovi dostanou, není správné je z analýzy odstranit – minimálně je potřeba tuhle informaci zmínit v analýze, např. uděláme analýzu se všemi hodnotami a poté další bez daných hodnot.

Reprezentativnost: náhodný výběr má obsahovat hodnoty ze všech podmínek procesů/skupiny. Například při anketě/ průzkumu by měl výběr zahrnovat zástupce:

- Ženského, mužského i nespécifikovaného pohlaví



- Různé výšky



- Různá povolání



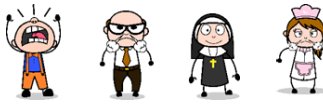
- Různé věkové skupiny



- Různé národnosti



- Různé nálady





Ve výrobním procesu to může být odebrání vzorků na základě definovaných pravidel:

- Časový úsek: každé 4H/ 3 kusy; 1× za týden 3 ks (zástupce výrobní rodiny);
- Stav výroby: 3 kusy na začátku a konci výroby; 3 kusy po přeseřízení výroby anebo po přestávce.

1.2 Různé druhy výběru

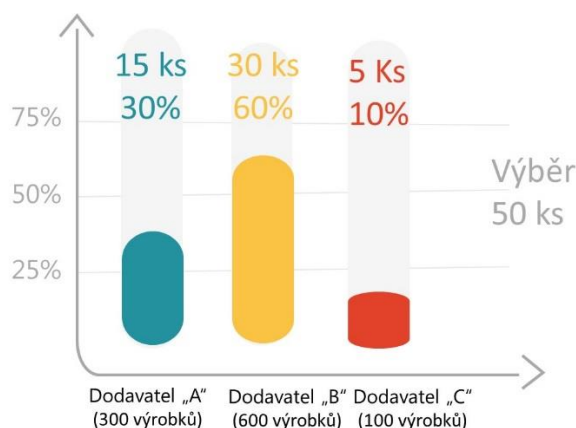
Jednoduchý náhodný výběr – procedura, při níž má každá položka stejnou pravděpodobnost, že bude vybraná (např. losování čísel při sportce/ bingo ...) [2].

Stratifikovaný (vrstvený) výběr rozdělíme vzorky do skupin a následně vybíráme poměrné zástupce z jednotlivých skupin. Například máme:

- 300 výrobků dodavatel „A“,
- 600 výrobků dodavatel „B“ a
- 100 výrobků dodavatel „C“.

Pokud budeme chtít vybrat 50 výrobků, náš výběr se bude skládat z:

- 15 výrobků dodavatele „A“ (30 %)
- 30 výrobků dodavatele „B“ (60 %) a
- 5 výrobků dodavatele „C“ (10 %) [2].



Systematický výběr je výběr na základě definovaných intervalů výběru vzorku z populace [3].

1.3 Vliv počtu vybraných kusů



Ted' si ukážeme vliv počtu vybraných kusů na výsledek analýzy.

Pro tenhle případ jsem vygeneroval 1.000 hodnot s normálním rozdělením (50; 0,25), které budou představovat všechny změřené kusy našeho procesu. Z téhle skupiny jsem následně vybral hodnoty, reprezentující náhodné měření výrobků ve výrobě:

- 100: změřili jsme každý 10. výrobek z procesu;
- 10×5: změřili jsme prvních 5 výrobků a po výrobě 100. výrobku jsme změřili dalších 5 kusů;
- 10: změřili jsme každý 100. výrobek z procesu.

Níže můžete vidět statistické parametry pro jednotlivé skupiny analyzovaných dat.

Variable	Mean	StDev	Minimum	Median	Maximum	Range
1.000	50,001	0,259	49,160	50,010	50,920	1,760
100	50,050	0,236	49,410	50,030	50,640	1,230
10×5	49,992	0,276	49,260	50,015	50,520	1,260
10	50,139	0,180	49,870	50,160	50,400	0,530

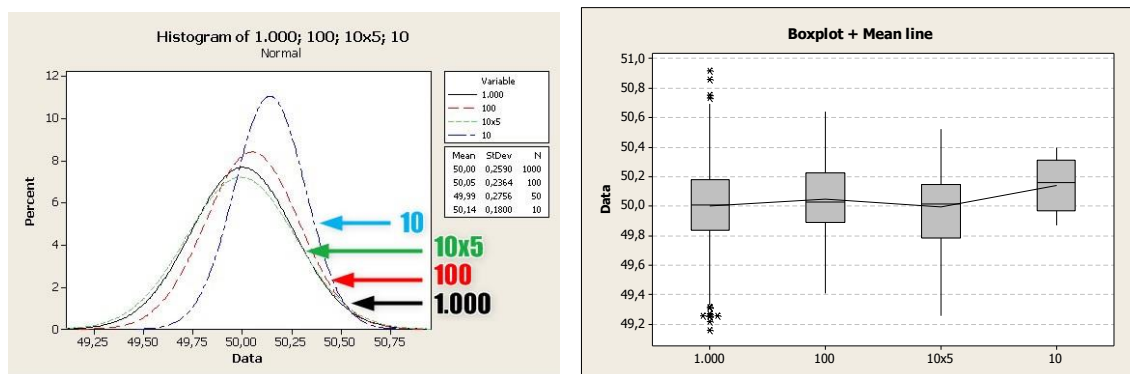
Například je možné vidět rozdíl v průměrné hodnotě (*Mean*) a směrodatné odchylce (*StDev*) – ve dvou základních parametrech, pomocí kterých se počítá „odhadovaná“ způsobilost a PPM procesů.

Pro naši demonstraci zanedbáme, že máme malý počet hodnot (učili mě, že statistické minimum je 50 hodnot a např. pro Ford je to 125 hodnot). Pokud dané hodnoty použijeme pro „odhad“ Cpk/Ppk a PPM našeho imaginárního procesu se specifikací 50 ± 1 , dostaneme výsledek níže:

Variable	Cpk	Ppk	PPM (Overall)
1.000	1,323	1,286	112,70
100	1,318	1,339	33,84
10×5	1,173	1,200	287,09
10	1,224	1,594	0,87

Ve skutečnosti náš proces vykazuje 112,7 PPM (1.000 hodnot), ale například při výběru každého 100. výrobku (10 hodnot) to bude jenom 0,87 PPM. Z daného důvodu se to nazývá „odhad“ způsobilosti procesů, pokud nepracujeme se všemi hodnotami, ale s výběrem.

Níže je grafické zobrazení našeho procesu a procesu na základě výběru.

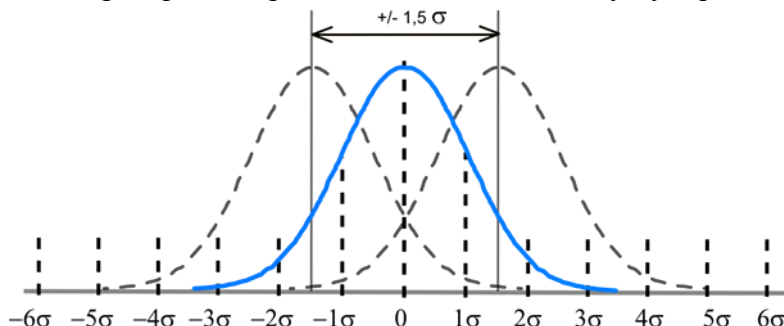


Daná prezentace je jenom ukázka hodnot vygenerovaných generátorem náhodných čísel s normálním rozdělením. Bohužel v reálném životě nemusí mít proces normální rozdělení a v reálné produkci má na výrobní proces vliv více faktorů jako například:

- Vstupní materiál, který má svou toleranci, různé výrobní šarže od dodavatele/-ů, kavita dílů ...;
- Opotřebení nástrojů a zařízení;
- Nastavení zařízení a jejich pravidelná údržba;
- Vliv obsluhy – vložení/ upevnění dílů do zařízení;
- Teplota / vlhkost ve výrobní hale v průběhu roku.



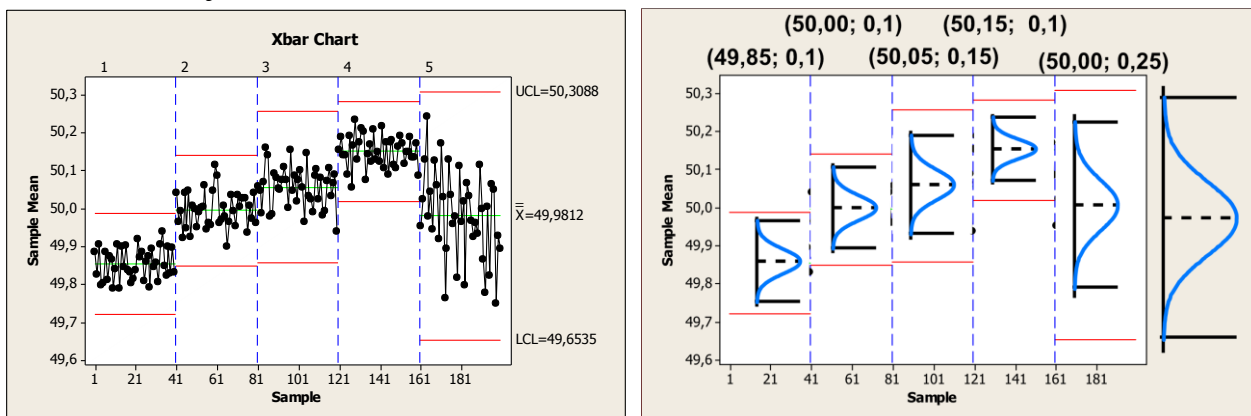
Například metodika Six Sigma počítá s posunem směrodatné odchylky v průběhu času $\pm 1,5\sigma$.



Pro další analýzu jsem nasimuloval změnu hodnot procesů. Opět jsem použil generátor náhodných čísel pro 5 výrobních dávek po 200 hodnot s parametrem (μ, σ):

- 49,85; 0,1 / 50; 0,1 / 50,05; 0,15 / 50,15; 0,1 / 50; 0,25

Na obrázku níže je znázorněné rozložení dat



Níže můžete vidět statistické parametry pro jednotlivé skupiny analyzovaných dat.

Variable	Mean	StDev	Minimum	Median	Maximum	Range
1000	50,007	0,179	49,280	50,010	50,570	1,290
100	50,002	0,183	49,650	49,980	50,480	0,830
10×5	50,023	0,145	49,710	50,030	50,280	0,570
10	50,091	0,110	49,950	50,075	50,300	0,350

Opět je možné vidět rozdíl v průměrné hodnotě (*Mean*) 1.000 hodnot = 50,007 a 10 hodnot = 50,091, stejně tak ve směrodatné odchylce (*StDev*).

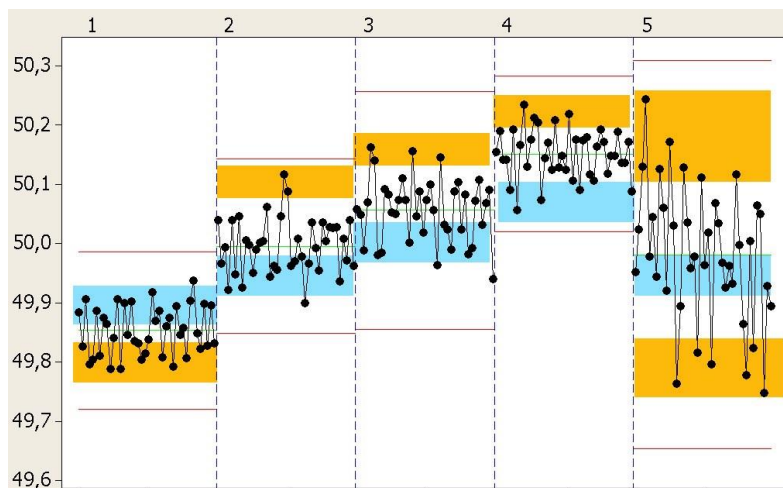
Pro výpočet způsobilosti a PPM procesů teď zpřísníme limit procesů na $50 \pm 0,7$ z důvodu zvýraznění rozdílů ve výsledku. Opět se objevily 2 situace zkreslující představu o našem procesu:

- Lepší výsledek: Při výběrů „10“ máme PPM = 0,01, ale ve skutečnosti má náš proces PPM = 90,04 a
- Horší výsledek: Při výběru „100“ máme PPM = 135,72, ale ve skutečnosti má náš proces PPM = 90,04

Variable	Cpk	Ppk	PPM (Overall)
1.000	1,678	1,293	90,04
100	1,488	1,268	135,72
10×5	1,899	1,556	1,83
10	1,662	1,849	0,01

Statistika je o náhodě:

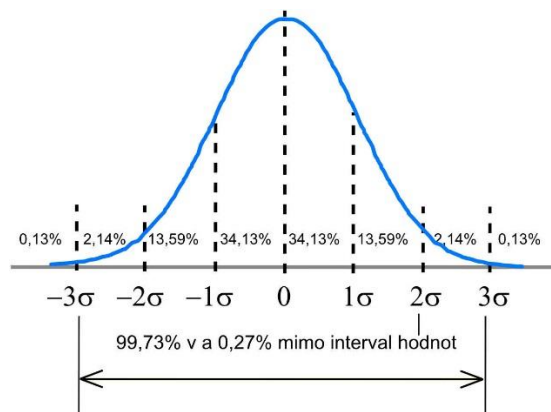
- Který výrobek z výběru (populace) vybere tzv. „šťastná ruka“ – povede se vybrat kus z modré zóny – menší rozsah hodnot procesů, anebo oranžové zóny – větší rozsah hodnot procesů (obr. níže)?
- Jak moc se jednotlivé výrobky ve výrobních šaržích liší (kombinace kavit, tolerance vstupních dílů ...).



A to samozřejmě nezahrnujeme vliv zkreslení výsledků zahrnutím jenom „určitých hodnot“ do analýzy. Jak mi někdo řekl, „*nevěřím statistice, kterou jsem si sám nefalšoval*“ anebo „*statistika je o prezentování předem zadaného/ specifikovaného výsledku*“ – náš proces je způsobilý / neobsahuje extrémní hodnoty = proto se může stát, že „někdo“ oranžové oblasti odstraní z analýzy; ale minimálně interně je nezbytné je analyzovat (vědět, jak proces ve skutečnosti vypadá) a zjistit, co se za nimi skrývá a co je způsobilo.

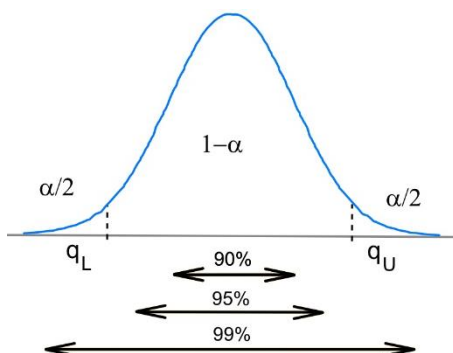
2 Bodový a intervalový odhad pravděpodobnosti

Jak už nejspíše víte, při normálním rozdělení má 99,73 % hodnot ležet v intervalu $\pm 3\sigma$ a 0,27 % hodnot mimo tento interval.



Možná jste se už setkali s pojmem riziko odhadu „ α “. V případě $\pm 3\sigma$ proces je $\alpha = 0,0027$ (0,27 %). Ve stručnosti jde o definování, s jak velkou nejistotou/přesností analýzy počítáme. Čím vyšší je hodnota nejistoty (riziko odhadu) α , tím nepřesnější je výsledek. Standardně se používá $\alpha = 0,05$ (5 %), tedy pravděpodobnost 95 %. Na obrázku níže je znázorněná oblast pro 99/95/90 %.

Riziko odhadu α je parametr, kterým je možné jednoduše zkreslit výsledek. V praxi to může být „odhad rizika“ NOK kusů v „poli“ / *Field* (provozu) při reklamaci. Pokud dáme $\alpha = 0,05$ (5 %), například dostaneme, že v „poli“ se může ještě najít 8 kusů s danou náhodnou chybou, ale při $\alpha = 0,1$ (10 %) to bude například jenom 1 kus (prezentujeme tedy nižší riziko pro zákazníka).



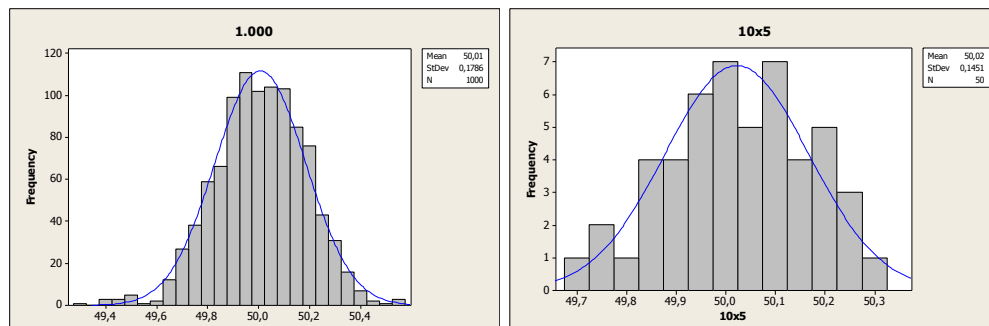
- $1-\alpha$: spolehlivostní odhad;
- α : riziko odhadu;
- q_L : dolní hranice intervalu;
- q_U : horní hranice intervalu.

Zadaná pravděpodobnost α udává pravděpodobnost vzniku ukazatele spolehlivosti v daném intervalu, a proto se obvykle nazývá spolehlivostní pravděpodobnost (konfidenční pravděpodobnost). Interval při zadané pravděpodobnosti se nazývá interval spolehlivosti (konfidenční interval) [4].

2.1 Intervalový odhad pravděpodobnosti

Pro další analýzu budu používat soubor dat s 5 šaržemi a 2 soubory dat:

- 1 000: (základní soubor) hodnoty všech výrobků a
- 10×5: (výběr) změřili jsme prvních 5 výrobků a po výrobě 100. výrobku změřeni dalších 5 kusů.

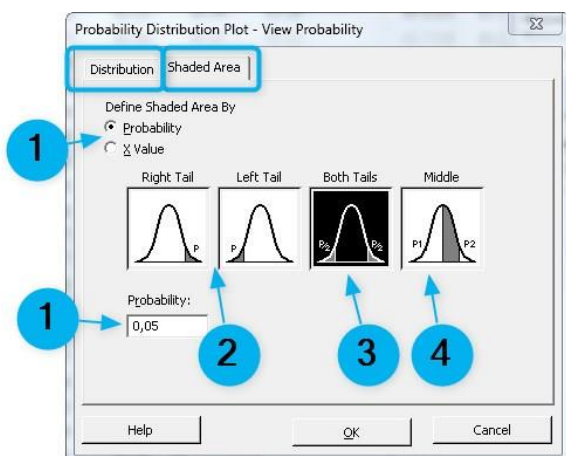


Pro další analýzu potřebujeme vědět o procesech jenom 2 statistické hodnoty – průměrnou hodnotu (μ /Mean) a střední hodnotu (σ /StDev).

Variable	Mean	StDev	Minimum	Median	Maximum	Range
1000	50,007	0,179	49,280	50,010	50,570	1,290
10×5	50,023	0,145	49,710	50,030	50,280	0,570

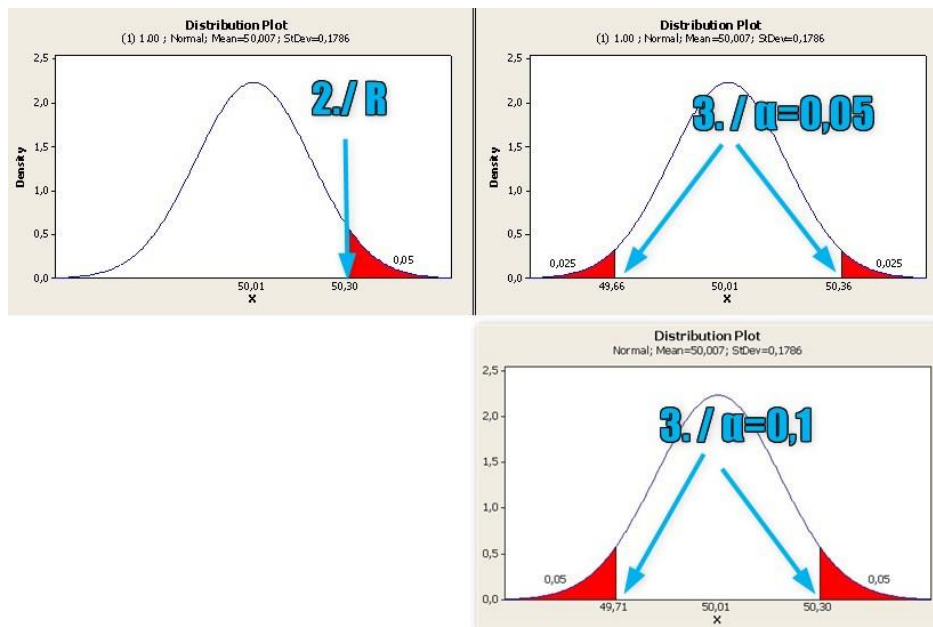
V MiniTab-u vybereme „**Graph** → **Probability Distribution Plot** → **View Probability**”

- V kartě „*Distribution*“ nastavíme parametry:
 - Distribution: Normal (normální rozdělení)
 - Men (střední hodnota) a Standrad deviation (směrodatná odchylka) procesů
- Pak v „*Shaded Area*“ máme možnost:
 1. Jestli chceme definovat hranici intervalu na základě pravděpodobnosti (*Probability*) anebo hodnoty (*X Value*) a následně s jakou hodnotou
 2. Zobrazení intervalu horní/pravý limit (*Right Tail*) anebo dolní/levý limit (*Left Tail*)
 3. Definované oba limity (*Both Tails*)
 4. Definované intervaly od-do (*Middle*)

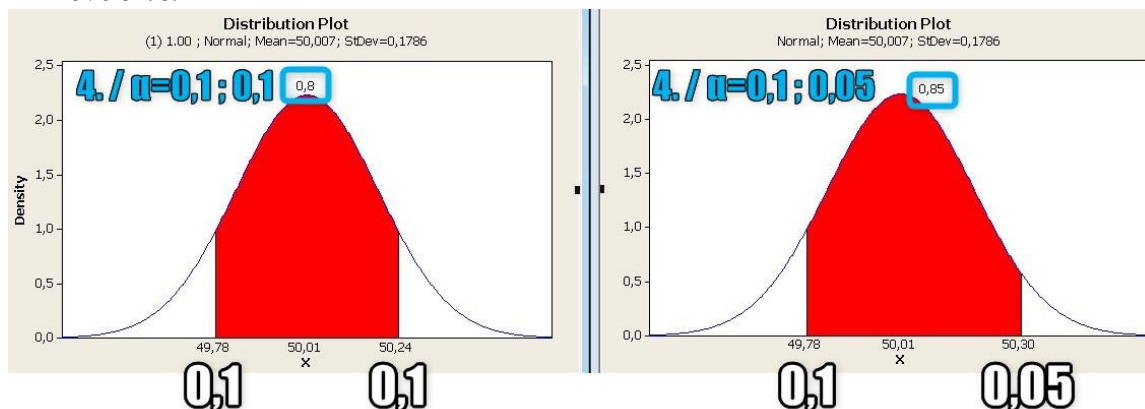


Ted' si ukážeme výsledek, pro soubor „1.000“ se zobrazením hodnoty na základě pravděpodobnosti.

- Chceme vědět limit pro 5 % (0,05) hodnot ležících nad horním limitem (2./ *Right*). Jedná se o hodnotu 50,3. Tedy 5 % hodnot bude mít hodnotu větší než 50,3.
- Pokud chceme oba limity (3. *Both*), při tomhle zobrazení se oblast rozdělí na $\frac{1}{2}$ (teda 2,5 % vlevo a 2,5 % vpravo při pravděpodobnosti 5 %). Pokud chceme 5 % vlevo i vpravo (spodní a horní limit), musíme zadat pravděpodobnost 10 % (0,1).



- Anebo (4) zobrazení oblasti (*Middle*) hodnot ležících pod křivkou histogramu definovanou limitem. 80 % hodnot bude ležet v rozsahu 49,78 až 50,24 (při pravděpodobnosti 10 % z pravé i levé strany). Další graf zobrazuje plochu 85 % se zobrazením z pravé strany 10 % a z levé 5 %.



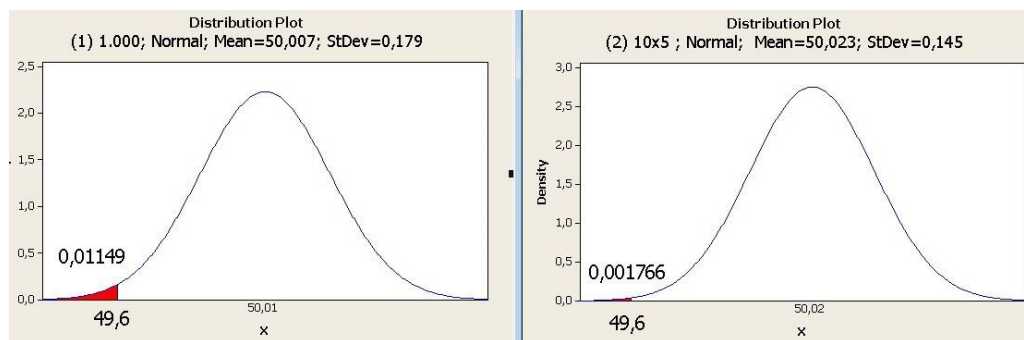
Podle mého názoru se v praxi spíše využije zobrazení na základě definování hodnoty. Tenhle případ si zobrazíme na obou souborech dat. PPM pro dané 2 soubory dat se specifikací procesů $50 \pm 0,7$ jsou:

Variable	PPM<LSL	PPM>USL	PPM Total
1000	37,47	52,57	90,04
10×5	0,31	1,52	1,83

Využití zobrazení procenta hodnot na základě uvedené hodnoty má v praxi využití pro výpočet, jaké procento hodnot našeho procesu by se „statisticky“ mělo nacházet mimo danou hodnotu.

Např. u našeho procesu s limitem $50 \pm 0,7$ (49,3–50,7) chceme vědět, jaké % hodnot je mimo specifikovaný limit (zkoumanou hodnotu), a z toho si můžeme odvodit PPM procesu. Níže v tabulce je zobrazen výsledek pro hodnotu 49,3 a 49,6 (chceme vědět, kolik hodnot bude pod zkoumanou hodnotou).

	Hodnota (spodní limit)	Pravděpodobnost (% hodnot)	PPM
1.000	49,3	0,00003912	39,12
	49,6	0,01149	11.490
10×5	49,3	3,0788E-07	0,307
	49,6	0,001766	1.766



To, co jsme si ukázali, byl jenom „statistický odhad“ na základě známé průměrné hodnoty a směrodatné odchylky procesů. Bohužel to nemusí odpovídat realitě. Pro zajímavost je v tabulce níže zobrazené skutečné PPM našeho procesu při zkoumaných limitech:

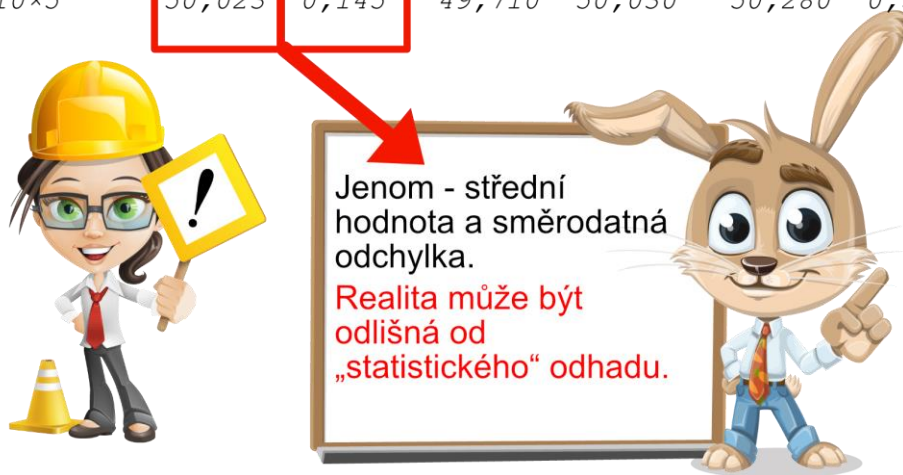
- 49,3: v našem souboru dat je 1 hodnota pod limitem, tedy 0,1 % hodnot a to je 1.000 PPM proti statistice, která odhadovala 39 PPM.
- 49,6: v našem souboru dat je 15 hodnot pod daným limitem, tedy 1,5 % hodnot a to je 15.000 PPM proti statistice, která odhadovala 11.490 PPM (což je rozdíl 23 %).

Pro soubor dat „10×5“ byl výsledek 0 PPM, v daném výběru dat nebyl žádný kus, ale víme, že kus mimo specifikovaný limit jsme v souboru všech dat měli – jenom jsme neměli to „štěstí“ ho vytáhnout. Další problém, který může nastat, je to, že náš proces nemá normální rozdělení.

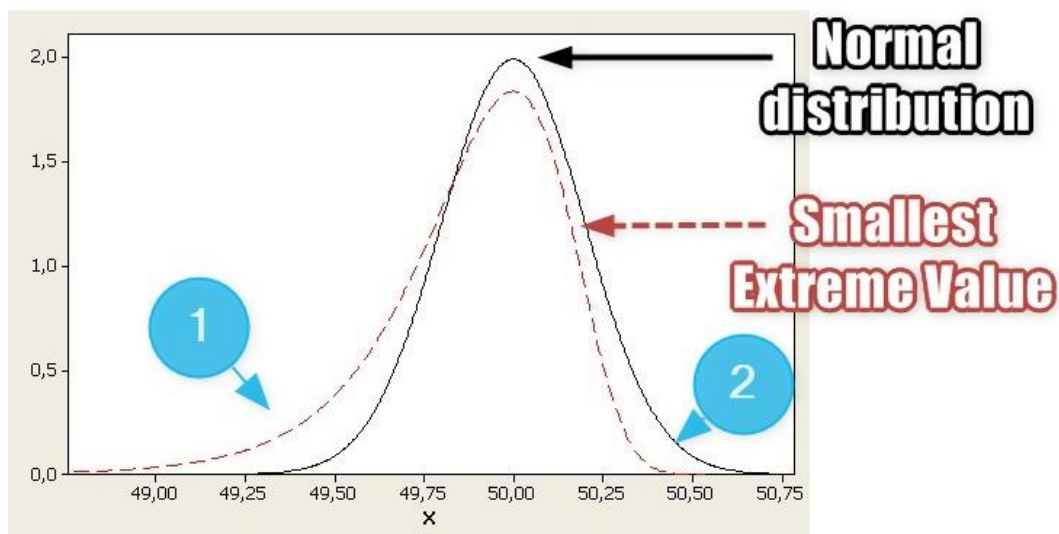
	Počet hodnot	Hodnota (spodní limit)	Počet hodnot pod limitem	Procento	PPM (skutečnost)	PPM (statistika)
1.000	1000	49,3	1	0,10 %	1.000	39,12
	1000	49,6	15	1,50 %	15.000	11.490
10×5	50	49,3	0	0,00 %	0	0,307
	50	49,6	0	0,00 %	0	1.766

Je potřebné si uvědomit, že při výpočtu způsobilosti a PPM procesu se berou v úvahu jenom 2 statistické parametry střední hodnota a směrodatná odchylka. Realita tedy může být odlišná od „statistického“ odhadu.

Variable	Mean	StDev	Minimum	Median	Maximum	Range
1000	50,007	0,179	49,280	50,010	50,570	1,290
10×5	50,023	0,145	49,710	50,030	50,280	0,570



Níže je ukázka porovnání Normálního rozdělení a Extrémně malé (nízké) hodnoty. Na rozdíl od normálního rozdělení můžeme vidět, že křivka je u spodního limitu „protáhlejší“ (1) = je tam vyšší pravděpodobnost hodnot než u normálního rozdělení. Naopak u vyšších hodnot je zkosenější (2) = je tam nižší pravděpodobnost hodnot u horního limitu.

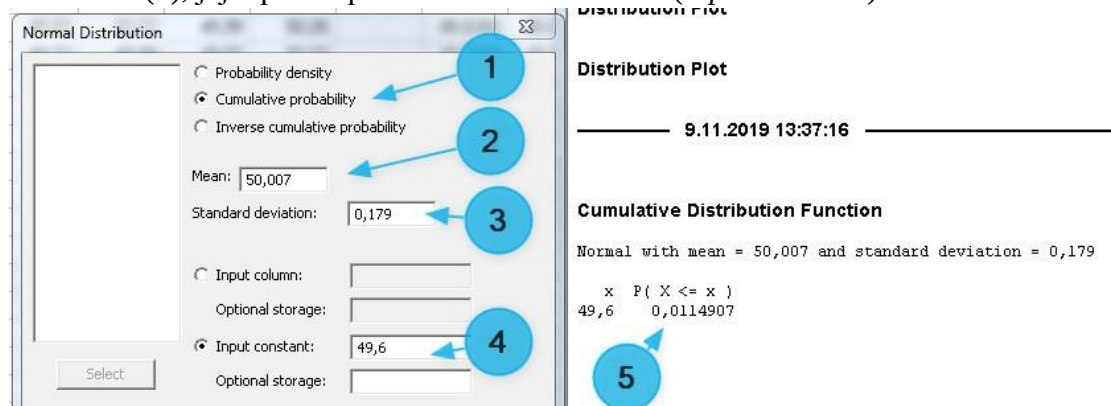


2.2 Bodový odhad - výpočet pravděpodobnosti konkrétní hodnoty

Pokud potřebujeme vypočítat pravděpodobnost určité hodnoty, poslouží k tomu funkce CDF (*Cumulative Distribution Function*). Pro normální rozdělení v programe MiniTab to je:

- **Calc** → **Probability Distributions** → **Normal**; následně vybrat „Cumulative probability“ (1)

Je potřeba zadat průměrnou hodnotu (*Mean*) (2), směrodatnou odchylku (*Standard deviation*)(3) a hodnotu (4), jejíž pravděpodobnost chceme vědět (*Input constant*).



Pro ukázkou jsem použil data pro základní soubor „1.000 hodnot“ a chtěl jsem znát pravděpodobnost hodnoty 49,6, což je 0,0114907 (tedy 1,14 %) (5).

V předešlé kapitole jsme si ukázali pravděpodobnost hodnot pod/nad/mezi určitou hodnotou, CDF je pravděpodobnost konkrétní hodnoty. Daná funkce by se dala použít například pro případ, kdy chceme znát pravděpodobnost konkrétní výšky lidí / velikosti nohy / věku ...

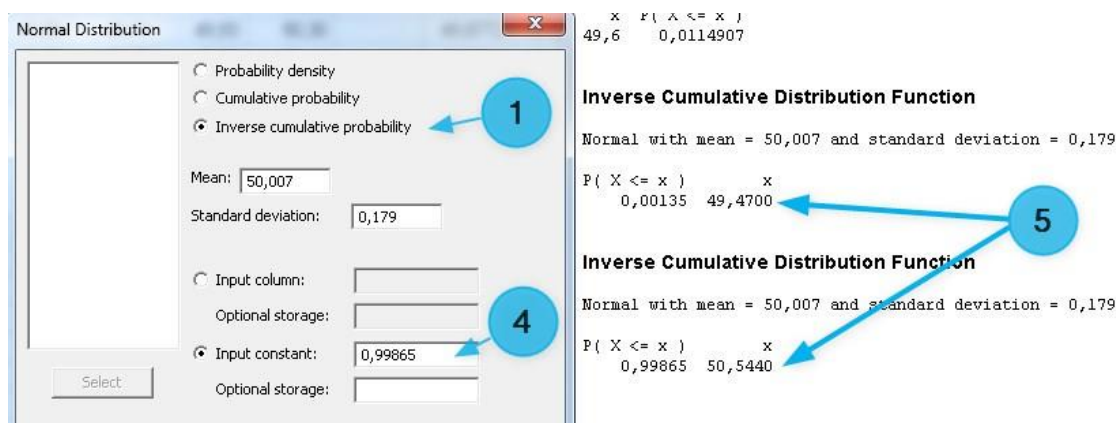
Pokud chceme znát hodnotu, která má definovanou pravděpodobnost, například chceme vědět hodnotu $z\ 3\sigma$, co je 99,73 %. Pomocný výpočet je:

$$1 - 0,9973 = 0,0027$$

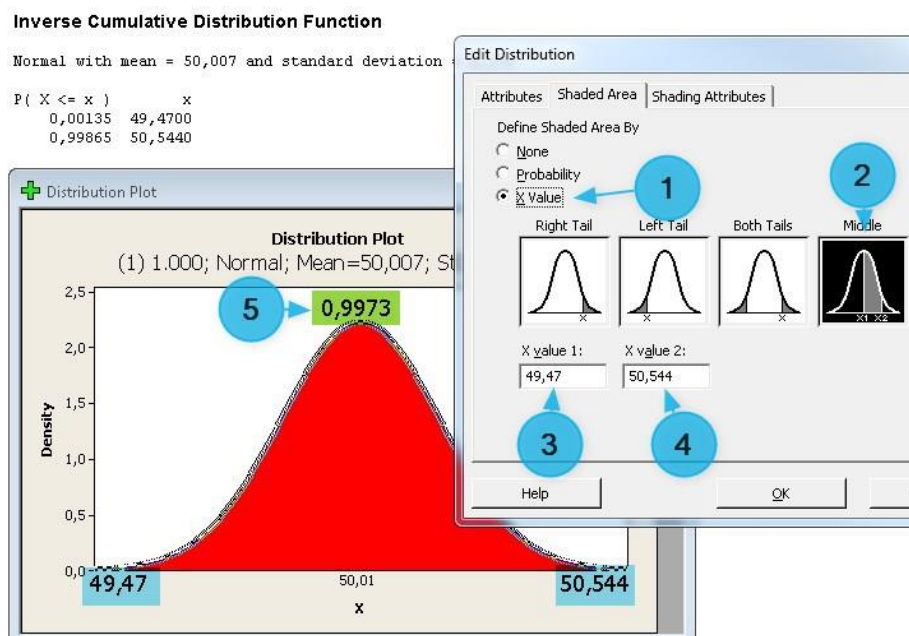
$$\frac{0,0027}{2} = 0,00135$$

$$1 - 0,00135 = 0,99865$$

Chceme znát hodnotu pravděpodobnosti 0,00135 (spodní limit) a 0,99865 (horní limit). K tomu slouží možnost „Inverse cumulative probability“.

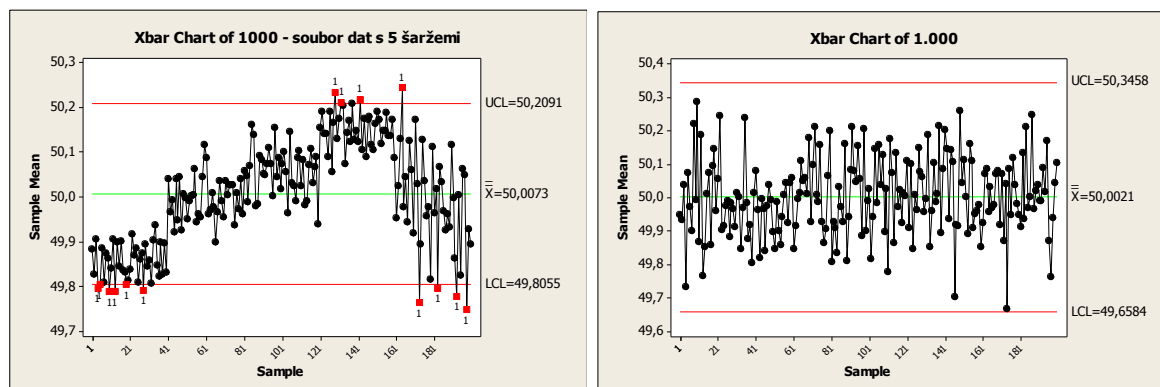


Je to hodnota 49,47 pro spodní limit a 50,544 pro horní limit. Mezi těmito dvěma hodnotami by mělo být 99,73 % hodnot našeho souboru dat.



Pokud bychom si dané hodnoty chtěli ověřit výpočtem způsobilosti, který by měl být $C_p/P_p=1$, narazíme na problém s tím, že náš soubor dat není zrovna „ideální“, ale na druhou stranu odráží to, s čím se můžete setkat v praxi – soubor dat obsahuje různé šarže materiálu, a proto je rozdíl v celkové (*Overall*) a potenciální (*Within*) směrodatné odchylce, jež má vliv na výpočet způsobilosti procesů a PPM.

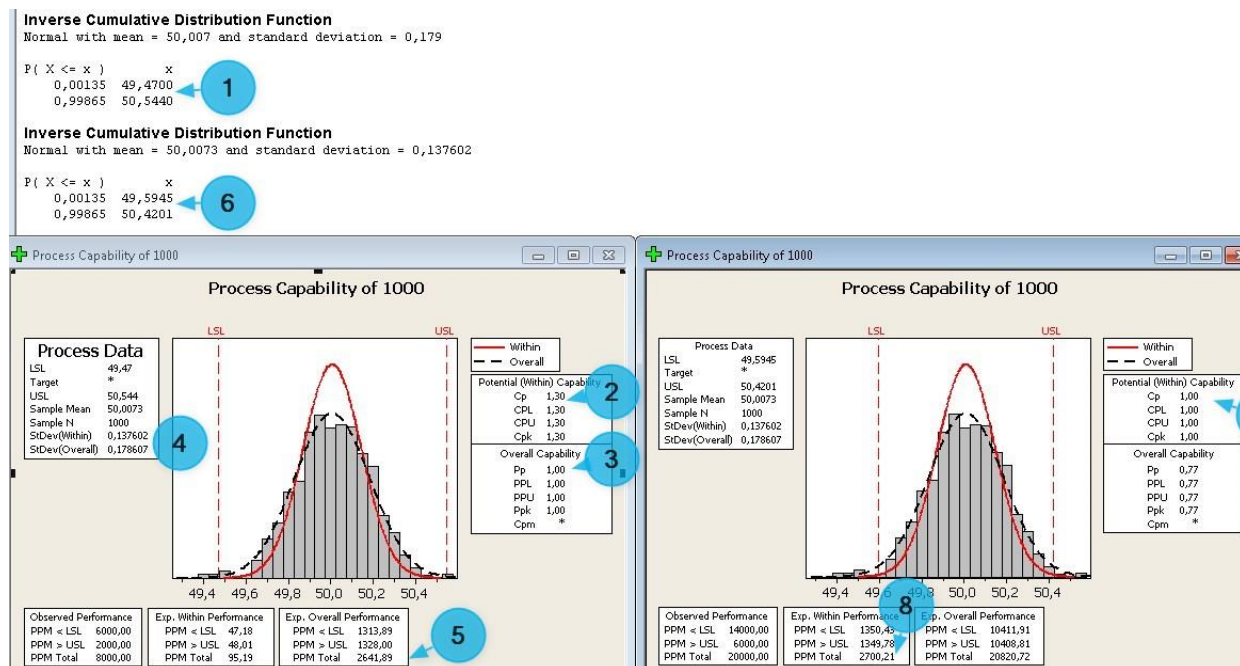
V našem případě se navíc jedná o extrémní případ, který je „statisticky nezvládnutý“, jak můžeme vidět na SPC grafu (minimálně některé hodnoty se nacházejí mimo regulační limity, dále je možné vidět trend). Vedle můžeme vidět SPC graf pro vygenerované hodnoty jako jeden balík.



Při zadání hodnot (1) jako limit získáme pro celkovou způsobilost (3) $P_p = 1$, ale (2) $C_{pk} = 1,3$ a (5) $PPM = 2641$ místo 2700.

Proto jsem vypočítal nové limity (6) se směrodatnou odchylkou (4) „0,137602 (*Within*)“. Pak (7) $C_p = 1$ a (8) $PPM = 2700$ (*Within*).

Podle mě je lepší postup první, kdy jsme použili celková data (*Overall*), ne potenciální výsledek (*Within*). Realita bývá horší, než doufáme, že se bude náš proces chovat.



Použitá literatura

- [1] Statistické metody řízení jakosti, ČSJ, Praha 2007, ISBN 978-80-02-01897-1, s. 23;
- [2] Connie M. BORROR, The Certified Quality Engineer, Third edition, American Society for Quality, Quality Press, Milwaukee 2009, ISBN 978-0-87389-745-7, s. 364
- [3] Issa BASS, Six Sigma Statistic with Exel and Minitab, The McGraw-Hill Companies, 2007, 0-07-154268-X; s. 100
- [4] Riadenie Kvality, Jozef HRUBEC, Slovenska poľnohospodárska univerzita v Nitre, 2000, ISBN 80-7137-849-6, s. 70